Human Reproduction, Vol.31, No.12 pp. 2714–2722, 2016
Advanced Access publication on September 22, 2016 doi:10.1093/humrep/dew227

human reproduction

REVIEW *Infertility*

No common denominator: a review of outcome measures in IVF RCTs

Jack Wilkinson[1,2,*], Stephen A. Roberts[1], Marian Showell[3], Daniel R. Brison[4], and Andy Vail[1,2]

[1]Centre for Biostatistics, Institute of Population Health, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 9PL, UK [2]Research & Development, Salford Royal NHS Foundation Trust, Salford M6 8HD, UK [3]Cochrane Gynaecology and Fertility, The University of Auckland, Auckland City Hospital, Auckland 1142, New Zealand [4]Department of Reproductive Medicine, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre (MAHSC), Manchester M13 9WL, UK

*Correspondence address. Centre for Biostatistics, Institute of Population Health, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 9PL, UK. E-mail: jack.wilkinson@manchester.ac.uk

Submitted on June 14, 2016; resubmitted on August 3, 2016; accepted on August 10, 2016

STUDY QUESTION: Which outcome measures are reported in RCTs for IVF?

SUMMARY ANSWER: Many combinations of numerator and denominator are in use, and are often employed in a manner that compromises the validity of the study.

WHAT IS KNOWN ALREADY: The choice of numerator and denominator governs the meaning, relevance and statistical integrity of a study's results. RCTs only provide reliable evidence when outcomes are assessed in the cohort of randomised participants, rather than in the subgroup of patients who completed treatment.

STUDY DESIGN, SIZE, DURATION: Review of outcome measures reported in 142 IVF RCTs published in 2013 or 2014.

PARTICIPANTS/MATERIALS, SETTING, METHODS: Trials were identified by searching the Cochrane Gynaecology and Fertility Specialised Register. English-language publications of RCTs reporting clinical or preclinical outcomes in peer-reviewed journals in the period 1 January 2013 to 31 December 2014 were eligible. Reported numerators and denominators were extracted. Where they were reported, we checked to see if live birth rates were calculated correctly using the entire randomised cohort or a later denominator.

MAIN RESULTS AND THE ROLE OF CHANCE: Over 800 combinations of numerator and denominator were identified (613 in no more than one study). No single outcome measure appeared in the majority of trials. Only 22 (43%) studies reporting live birth presented a calculation including all randomised participants or only excluding protocol violators. A variety of definitions were used for key clinical numerators: for example, a consensus regarding what should constitute an ongoing pregnancy does not appear to exist at present.

LIMITATIONS, REASONS FOR CAUTION: Several of the included articles may have been secondary publications. Our categorisation scheme was essentially arbitrary, so the frequencies we present should be interpreted with this in mind. The analysis of live birth denominators was post hoc.

WIDER IMPLICATIONS OF THE FINDINGS: There is massive diversity in numerator and denominator selection in IVF trials due to its multistage nature, and this causes methodological frailty in the evidence base. The twin spectres of outcome reporting bias and analysis of non-randomised comparisons do not appear to be widely recognised. Initiatives to standardise outcome reporting, such as requiring all effectiveness studies to report live birth or cumulative live birth, are welcome. However, there is a need to recognise that early outcomes of treatment, such as stimulation response or embryo quality, may be appropriate choices of primary outcome for early phase studies.

STUDY FUNDING/COMPETING INTERESTS: J.W. is funded by a Doctoral Research Fellowship from the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. J.W. also declares that publishing research is beneficial to his career. J.W. and A.V. are statistical editors, and M.S. is Information Specialist, for the Cochrane Gynaecology and Fertility Group, although the views

expressed here are not necessarily those of the group. D.R.B. is funded by the NHS as Scientific Director of a clinical IVF service. The authors declare no other conflicts of interest.

Key words: IVF / outcome measures / assisted reproduction / core outcomes / live birth / IMPRINT / CROWN / infertility trial / ongoing pregnancy / reporting guidelines

Introduction

Inconsistency and incompleteness of outcome reporting in infertility trials are barriers to understanding and improving treatments (Dapuzzo *et al.*, 2011; Legro *et al.*, 2014). In the absence of common standards of reporting, it may be difficult to compare the safety and effectiveness of competing interventions, or to synthesise the results of trials in meta-analysis (Blazeby *et al.*, 2012; Khan, 2014; Clarke and Williamson, 2016). The choice of outcome also has implications for both the relevance (Heijnen *et al.*, 2004; Min *et al.*, 2004; Legro *et al.*, 2014) and methodological validity (Vail and Gardener, 2003; Griesinger, 2016) of a trial's results.

Choosing an outcome for trials of IVF is particularly complex, owing to the multistage nature of the treatment. Treatment comprises stimulation of the ovaries, retrieval and fertilisation of oocytes and the culture and transfer of some of the resulting embryos to the uterine cavity (Van Voorhis, 2007). Some of these embryos may implant, some of these may result in a clinical pregnancy, and some of these may result in a live birth. Those embryos not used for the initial transfer may be cryopreserved, so that they can later be thawed and transferred in a subsequent attempt. The response at each stage can be quantified: ovarian response by the number and maturity of oocytes; fertilisation by the number of zygotes, and subsequently the number and quality of embryos produced; the transfer procedure by the implantation of embryos; and the clinical outcome of treatment by clinical pregnancy and the birth of a child. Additionally, treatment may fail at each stage: stimulation may be cancelled due to poor or over-response; fertilisation failure may occur; embryos may fail to develop, or post-transfer fail to implant; and pregnancies may be lost before or subsequent to identification of a clinical pregnancy. One consequence of this for clinical trials of interventions designed to improve IVF is that numerous clinical and procedural events that occur during treatment can be reported. A second consequence is that these events may be reported in subgroups containing only those patients who reach a certain milestone, such as oocyte retrieval or embryo transfer. Further complexity arises due to the fact that IVF involves two or more individuals (e.g. a male and female partner), who may undertake multiple treatment cycles, and one or more additional individuals (babies) arising from successful treatment (Legro *et al.*, 2014). When selecting which outcomes to report in an IVF trial therefore, many numerators and denominators are available (Heijnen *et al.*, 2004).

The importance of the choice of numerator is well recognised and has been enshrined in the Improving the Reporting of Clinical Trials of Infertility Treatments (IMPRINT) statement with a call for live birth to be reported in all infertility trials (Legro *et al.*, 2014), although alternatives, such as ongoing pregnancy, have been proposed on pragmatic grounds (Braakhekke *et al.*, 2014a). The appropriate choice of denominator is a more subtle issue. The optimal denominator for IVF evaluation has been widely discussed (Germond *et al.*, 2004; Heijnen *et al.*, 2004; Abdalla *et al.*, 2010; Garrido *et al.*, 2011), and is known to have implications for the interpretation of trials, where the exclusion of randomised participants may introduce bias to the estimated treatment effect (Montori and Guyatt, 2001; Vail and Gardener, 2003; Mastenbroek *et al.*, 2005; Mastenbroek and Repping, 2014). This could occur if, for example, participants are randomised at the start of ovarian stimulation, but the outcome is calculated only in those who undergo transfer.

We conducted a review of outcomes reported in IVF RCTs in 2013 and 2014. Our aims were to establish the full range of outcomes in use in IVF RCTs and to identify the ramifications for the evidence base.

Materials and methods

Search strategy

M.S. performed a search of the Cochrane Gynaecology and Fertility Group PROCITE database on 22 June 2015 using the search strategy contained in Supplementary Data I. This is a specialised register of RCTs updated weekly by searching databases, conference abstracts and journals. Further details of the database are provided in Supplementary Data I. Our initial search covered the period 2010–2014, although we subsequently narrowed our focus to the period 2013–2014 owing to feasibility constraints. We screened the titles and abstracts of the identified articles and excluded those not meeting the eligibility criteria. We reviewed the full text of all articles not excluded during this initial screening phase and made further exclusions as appropriate.

Eligibility criteria

English-language publications of RCTs in peer-reviewed journals in the period 1 January 2013 to 31 December 2014 were considered eligible. Conference papers were excluded. We did not consider methodological quality to be relevant, as our concerns related to the outcomes reported in this literature and not in the estimation of treatment effects. To be eligible, a study had to have had participants undergoing IVF or ICSI including a period of ovarian stimulation in at least one arm of the trial, or participants undergoing frozen embryo transfer in at least one arm of the trial, or partners of patients undergoing IVF or ICSI in at least one arm of the trial, or oocyte donors donating to an IVF programme. We included trials where surplus oocytes had been obtained as part of IVF or ICSI treatment and an intervention was applied to these oocytes even if there was no intention to subsequently transfer any of the resulting embryos. Finally, the publication had to report clinical or preclinical outcomes to be eligible (which would exclude, for example, purely economic evaluations of interventions).

Data extraction

Initially, we performed a small pilot extraction of five reports to inform the extraction process used in the full sample, including the variables to be extracted and the formatting of this information. We extracted information at both study-level and at the level of each reported outcome in a

study. We defined an outcome as any post-randomisation variable presented separately for each arm in the study or as a comparison between study arms and recorded both the numerator and denominator used in the calculation. We did not record a reported outcome multiple times if it was presented for each of several subgroups, unless these were defined by excluding patients who did not reach a certain stage in the process. We also did not record outcomes multiple times where these corresponded to repeated measurements at several time points. At the study-level, we extracted details of the intervention and the stage in the treatment process at which the intervention was applied (pre-stimulation phase, stimulation phase, post stimulation including culture and selection of embryos, transfer, frozen transfer or intervention targeted at the male partner, such as manipulation or selection of sperm prior to ICSI). Similarly, we extracted the stage of treatment at which randomisation took place. For each reported outcome, we extracted the numerator and denominator (for numerical variables, the denominator would be the divisor used in the calculation of a mean). Where pregnancy or live birth were reported, we extracted the corresponding definition used by the study authors. Data were extracted into two databases, one containing study-level information and another containing reported-outcome-level information. J.W. performed data extraction for all studies. S.R. and A.V. performed double extraction for a random sample of 10%, to check data quality and consistency of recording. Furthermore, we conducted extensive data validation and cleaning, including manually checking every entered item.

Statistical analysis

We summarised the characteristics of the sample and tabulated the numerators and denominators in use in nine categories (live birth, pregnancy, stimulation response, transfer, fertilisation, multiple births or pregnancies, other preclinical outcomes, adverse events and post-natal). These categories are arbitrary and have been selected to facilitate the presentation of our results. We note here however that, since our analyses are descriptive and these categories are purely presentational, it would not affect our results were an outcome measure to be reported under one heading rather than another. Due to the large number of outcomes identified, we reported only those appearing in more than one study. We simplified the results by combining similar numerators and denominators. For example, we combined live birth with take home baby rate, and combined the denominators 'per patient with sufficient embryos' and 'per patient with sufficient blastocysts', where 'sufficiency' could be defined on the basis of quantity or quality of embryos (or both). For this primary analysis, we did not distinguish between subtly different definitions of outcomes (e.g. clinical pregnancy may have been defined as foetal heartbeat on ultrasound at different time points in different studies). However, at the suggestion of an anonymous peer reviewer, we also present the definitions used by trial authors for pregnancy and live birth outcomes. In order to investigate the methodological implications of denominator selection, we conducted post-hoc analysis in the subgroup of studies reporting live birth. We recorded whether the denominator used coincided with the cohort of randomised participants (ignoring exclusions due to protocol violations) and if not, the nature and extent of the exclusion. We did not perform statistical inference, because we have attempted to summarise all trials within the time period and it is not clear that inference would be meaningful.

Sample size

The decision to include all studies in the period 1 January 2013 to 31 December 2014 was made primarily on pragmatic grounds, on the basis that this would be sufficient to assess current practices in outcome reporting while proving to be feasible. A post-hoc calculation can be made however. A sample of size 142 yields a 76% probability of observing a relatively rare outcome (appearing in 1 out of every 100 studies) at least once.

Ethical approval

Ethical approval was not required as the study involved only the review of published research.

Results

The dataset used for analysis can be accessed in Supplementary Data II.

Results of the search

Fig. 1 shows the results of the search and screening process. The search identified 640 references published between 2013 and 2014. Following title and abstract screening, 488 references were discarded without further assessment. The remaining 152 articles were assessed further by reviewing the full texts and a further 10 were excluded for the reasons shown in Fig. 1. Finally, 142 RCTs were included in the analysis. Agreement between raters was almost universal, with one reviewer erroneously extracting one additional outcome from one study due to misreading the text.

Stage of intervention and randomisation

Interventions were delivered prior to ovarian stimulation in 20 (14%) articles, during stimulation in 51 (36%), post stimulation or during culture of embryos in 31 (22%), post culture but preceding transfer of embryos in 19 (13%) and following the transfer procedure in 3 (2%). Five (4%) were trials of interventions targeted at male partners and 13 (9%) featured interventions designed to improve outcomes after the vitrification and warming of oocytes or embryos. Randomisation occurred prior to stimulation in 62 (44%) articles, during stimulation in 17 (12%), post stimulation or during culture in 27 (19%) and post culture but prior to transfer in 23 (16%). The point of randomisation was unclear in 13 (9%) articles.

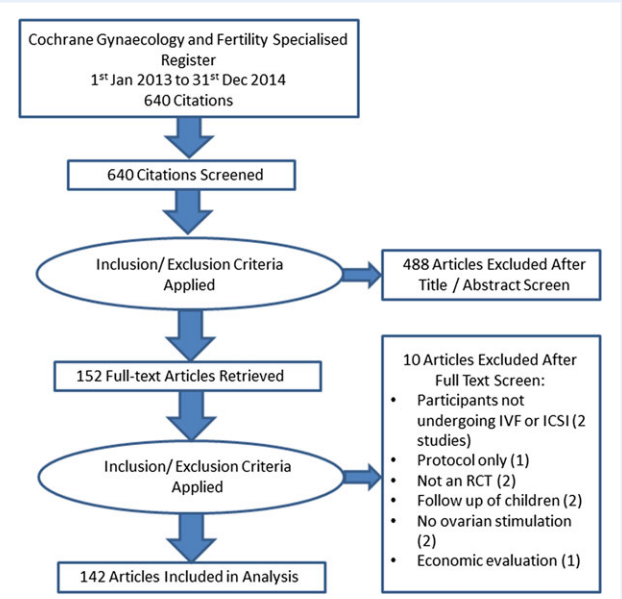

Figure 1 PRISMA diagram showing flow of studies in the review.

Reported outcomes

After combining similar items, 361 numerators and 87 denominators were discerned. A total of 815 distinct combinations of numerator and denominator were identified and 203 combinations appeared in more than one study (612 did not). The median (interquartile range: IQR) of distinct outcomes reported in a study was 11 (7–16), with a range of 1–36.

Live birth outcomes

Fifty-two (37%) articles reported the numerators live birth event or take home baby in total, with 14 combinations of numerator and denominator. Fig. 2 and Supplementary Table I show combinations of live birth numerators and denominators appearing in more than one study. It was most common to report these per transfer (15% of studies). Only 8 (6%) studies reported live birth per cycle started. It was not common (5%) for studies to report live birth in a cumulative fashion, across multiple fresh and frozen transfer cycles. No study reported cumulative live birth following multiple egg collections. Four (3%) reported cumulative live birth per cycle started and 2 (1%) reported time to pregnancy leading to live birth, where time was measured across multiple treatment cycles. Four (3%) studies reported a preterm birth event with three of these reporting preterm birth per baby.

Of the 52 studies reporting live birth rates, 22 (42%) used the point of randomisation as the denominator in the calculation. One study acknowledged that the calculation was not based on a randomised comparison and was therefore 'descriptive'. In six (12%) studies, the denominator could not be discerned. The remaining 23 (44%) did not use the randomised cohort as the denominator. In 17 (33%) studies, the denominator included only those undergoing transfer (15 studies) or oocyte retrieval (two studies) rather than the randomised participant. In these 17 studies, a median (IQR) of 8% (4–14%) of participants were excluded, with a range of (2–38%). Seven (13%) studies made a unit of analysis error when calculating live birth rates, with six calculating live birth rates per embryo transferred. In one trial each woman's oocytes were randomly split between intervention arms, and live birth per transfer was calculated in the subset of procedures where all embryos transferred had originated from one of the arms.

Pregnancy outcomes

Table I shows pregnancy outcomes appearing in more than one study. Forty-six (32%) reported biochemical pregnancy, with 13 different denominators. It was most common (16%) to report these per transfer procedure. Clinical pregnancy rates (with varying definitions) were reported in most (67%) studies, with 19 different denominators. Again, it was most common to report these per transfer procedure (31%) although the denominator 'per cycle started' was also reasonably prevalent (17%). Thirty-nine (27%) studies reported ongoing pregnancy using 16 different denominators, with 15% reporting ongoing pregnancy per transfer procedure. Only 5% reported ongoing pregnancy per cycle started. Very few studies reported clinical pregnancy (1%) or ongoing pregnancy (2%) in a cumulative fashion. Just under half (43%) reported miscarriages in addition to 6% reporting pregnancies that did not progress beyond the biochemical stage. Nineteen (13%) reported miscarriages per clinical pregnancy and 11 (8%) reported these per biochemical pregnancy.

Stimulation outcomes

Supplementary Table II shows outcomes relating to stimulation response. Number of oocytes (46%), of mature oocytes (23%), total gonadotrophin dose (27%) and stimulation duration (26%) were all commonly reported, each with a variety of denominators. Perhaps unsurprisingly, stimulation outcomes were more frequently reported per cycle started compared to pregnancy and live birth events; 28 (20%) reported number of oocytes, 17 (12%) reported number of mature oocytes, 21 (15%) reported gonadotrophin dose and 19 (13%) reported stimulation duration per cycle started. However, some studies did report stimulation outcomes in the subset of patients reaching later stages in the process (Supplementary Table II). Eighteen (13%) studies reported cycle cancellation, 13 (9%) per cycle started.

Fertilisation outcomes

Supplementary Table III shows fertilisation outcomes. Fertilisation (37%), the attainment of good quality embryos as a binary variable (15%), the number of embryos (19%), of good quality embryos (12%) and of frozen embryos (14%) were all frequently reported, each with a variety of denominators (Supplementary Table III). Other than cleavage (11%), no other numerator was reported in more than 8% of studies.

Transfer outcomes

Supplementary Table IV displays outcomes relating to the transfer procedure. Number of embryos transferred (52%) and implantation (52%) were the most commonly reported numerators in the review. The denominator used with implantation was often unclear (38%) but was otherwise generally reported per embryo transferred (30%) rather than as a patient-level outcome. Number of embryos transferred was most commonly reported per transfer procedure (17%). Other transfer outcomes appeared in relatively small numbers of studies; the next most recurrent was achievement of transfer (8%), reported per cycle started (4%) or per oocyte retrieval (2%).

Multiple pregnancies and births

Relatively few studies reported multiple pregnancies or births or pregnancies (Supplementary Table V). Seventeen percent of studies reported the numerator multiple pregnancy and 4% reported multiple birth rates. One study reported multiple pregnancy per cycle started, the only instance of an outcome in this category being reported with this denominator. Where multiple pregnancy was reported, it was not uncommon for it to be presented per clinical pregnancy (5%). Multiple birth was only reported per live birth event (3%) or per transfer (1%).

Other adverse events

The most commonly reported adverse event was ovarian hyperstimulation syndrome (OHSS) of unspecified severity (17%), with several studies specifying the severity as mild (3%), moderate (4%) or severe (4%) (Supplementary Table VI). Ectopic pregnancy rates were explicitly reported in 13% of studies and general adverse events were described in 6%.

Post-natal outcomes

Small numbers of studies reported post-natal outcomes, most commonly birthweight (6%), congenital abnormalities (4%) and gestational age (2%) (Supplementary Table VII). These were most frequently reported per baby.

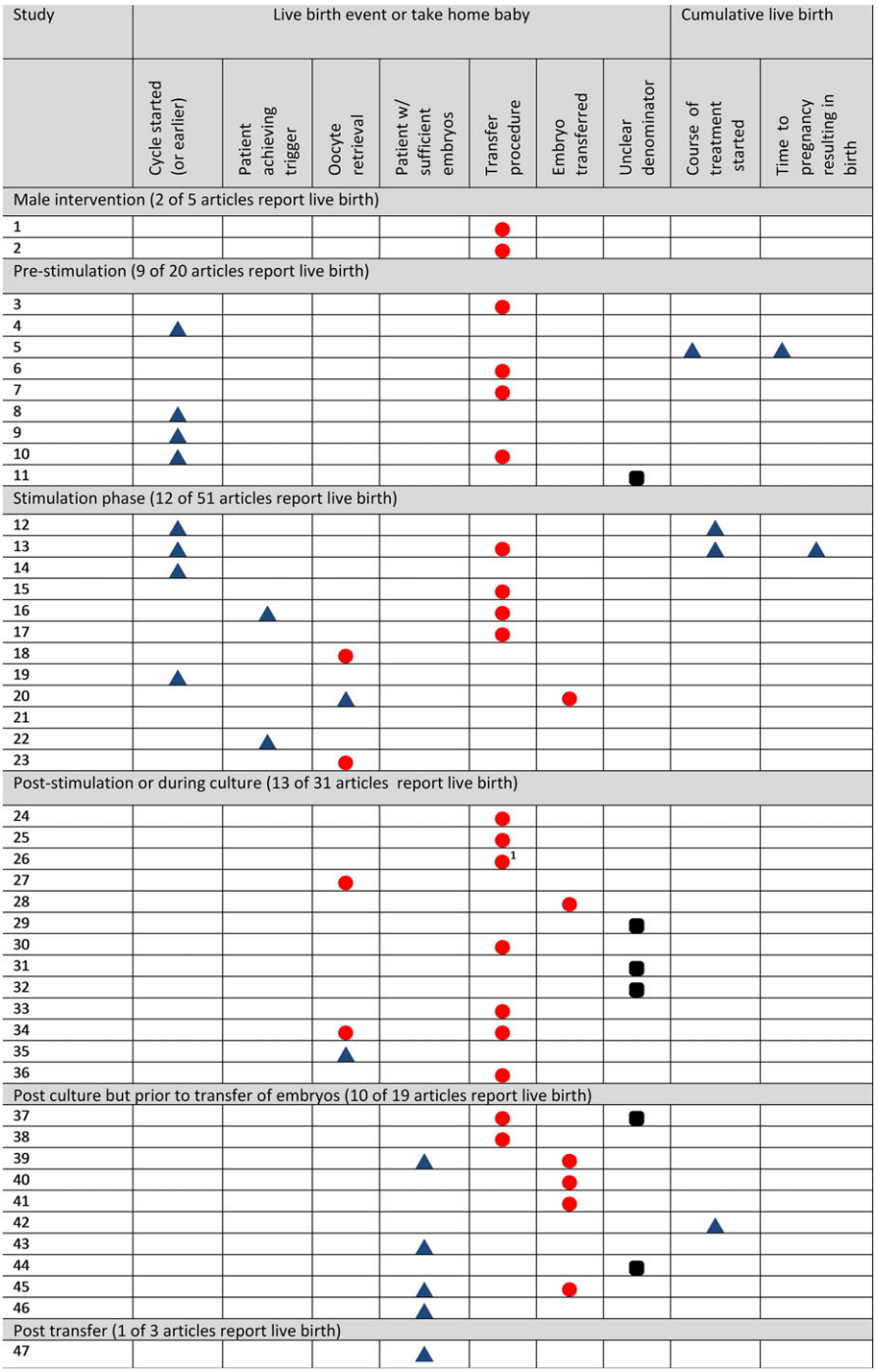

Figure 2 Reported live birth outcomes in 47 IVF RCTs in 2013–2014 by stage of intervention. Each row corresponds to a single study. Only studies reporting live birth outcome measures appearing in more than one study are shown (as opposed to all studies reporting live birth). Blue triangles (▲) indicate that the study authors used a denominator that coincided with the point of randomisation in the trial. Red circles (●) indicate that the study authors did not use the point of randomisation as the denominator, but instead included only patients who reached a certain stage of treatment when calculating live birth rates, potentially undermining the random allocation in the study. Black squircles (■) indicate that it is unclear whether or not the denominator coincided with the point of randomisation. [1]Authors presented this as a descriptive result.

Other procedural outcomes

Other procedural measurements were reasonably prevalent, such as estradiol levels (32%), endometrial thickness (25%) or progesterone levels (12%) (Supplementary Table VIII). These outcomes were generally reported using denominators including patients in the earlier stages of treatment (e.g. per cycle started or per oocyte retrieval).

Table I Pregnancy outcomes reported by more than one RCT. Frequency (%) of studies reporting each outcome (of a total *n* = 142).

Outcome numerator	Denominator	No (%) of studies reporting item
Biochemical pregnancy		46 (32% of studies)
	Per cycle started (or earlier)	12 (8%)
	Per transfer	23 (16%)
	Per patient achieving trigger	2 (1%)
	Per oocyte retrieval	2 (1%)
	Par patient w/sufficient embryos	5 (4%)
	Unclear denominator	2 (1%)
Biochemical pregnancy only		9 (6% of studies)
	Per transfer	2 (1%)
	Per transfer of embryos from one intervention arm only	2 (1%)
	Per chemical pregnancy	2 (1%)
	Unclear	2 (1%)
Clinical pregnancy		95 (67% of studies)
	Per cycle started (or earlier)	24 (17%)
	Per trigger	4 (3%)
	Per oocyte retrieval	11 (8%)
	Per patient w/sufficient embryos	6 (4%)
	Per transfer	44 (31%)
	Per transfer of embryos from one intervention arm only	3 (2%)
	Unclear	7 (5%)
Ongoing pregnancy		39 (27%)
	Per cycle started (or earlier)	7 (5%)
	Per oocyte retrieval	5 (4%)
	Per patient with sufficient embryos	5 (4%)
	Per transfer	21 (15%)
	Per clinical pregnancy	3 (2%)
Pregnancy (unclear)		9 (6%)
	Per cycle started (or earlier)	2 (1%)
	Per transfer	4 (3%)
Cumulative clinical pregnancy		2 (1%)
	Per course of treatment started	2 (1%)
Cumulative ongoing pregnancy		3 (2%)
	Per course of treatment started	2 (1%)
Miscarriage		61 (43%)
	Per chemical pregnancy	11 (8%)
	Per clinical pregnancy	19 (13%)
	Per cycle started (or earlier)	3 (2%)
	Per oocyte retrieval	3 (2%)
	Per transfer	9 (6%)
	Per transfer of embryos from one intervention arm only	2 (1%)
	Unclear	9 (6%)

Table II Frequency (%) of definitions of 'live birth' in IVF RCTs reporting on this outcome in 2013–2014.

Definition of live birth	Frequency (%) of studies
Birth of ≥1 neonate 28 weeks or later	1 (2)
Individual baby born after 24 weeks of gestation	2 (4)
Individual viable foetus at 24 weeks of gestation	1 (2)
Live birth event/delivery	19 (36)
Live birth event and individual baby (both given in article)	1 (2)
Individual living baby	1 (2)
Pregnancy >28 weeks of gestation	1 (2)
Undefined	27 (51)

Definitions of pregnancy and live birth used in the studies

Note that for these analyses, we have included the definitions used when variants of these outcome measures were reported, for example giving the definition of live birth used when cumulative live birth was reported. Accordingly, the totals for these analyses do not match those in the analyses described above.

Live birth

Table II shows the definitions provided by authors reporting live birth. It was most common (27 studies, 51%) for no definition to be given, followed by 19 (36%) defining this as a count of live birth events/deliveries. Other definitions, such as counts of individual babies, were sparse.

Clinical pregnancy

Supplementary Table IX shows the definitions of clinical pregnancy. This was not defined in around one fifth (*n* = 21, 21%) of studies reporting clinical pregnancy. A variety of subtly different definitions were used, with the vast majority comprising some combination of ultrasound confirmation of gestational sacs and foetal heartbeat at different time points.

Ongoing pregnancy

Supplementary Table X shows the definitions of ongoing pregnancy, with around a third (13 studies, 33%) of studies reporting this not providing any. Definitions were somewhat variable, with considerable differences in the gestational age required to declare that the pregnancy was ongoing.

Biochemical pregnancy

Supplementary Table XI shows the definitions of biochemical pregnancy. These were almost universally defined on the basis of positive B-hCG tests, with variations arising from different cut-off values of the assay and different timings of testing.

Discussion

Our review confirms large-scale diversity in outcome reporting in IVF trials and suggests several areas of systematic methodological weakness in the evidence base. Over 800 combinations of numerator and denominator were reported, the majority of which were not used in more than one article. No single outcome measure appeared in a majority of studies. Subtly different definitions of numerators were employed, increasing the variety of reporting options even further. This affirms the concerns highlighted by the Core Outcomes in Women's Health (CROWN) Initiative who noted that a lack of common reporting standards was a hindrance to the synthesis of evidence (Khan, 2014). The recommendation set out in IMPRINT, that all infertility trials should report live birth and cumulative live birth, may go some way to address this matter. This review indicates that at present a minority of studies report live birth and few report cumulative live birth, although it was not common for studies to include multiple treatment cycles. The rates of reporting of live birth and other clinical outcomes are lower than was observed in a previous review of infertility trials, because the authors of that study required the reporting of a clinical outcome for inclusion (Dapuzzo *et al.*, 2011). Moreover, we have shown that where live birth is reported, a variety of denominators are used. Consequently, we suggest that the matter of combining outcomes with different denominators in meta-analysis warrants attention. We note that the proposition to have live birth as the primary outcome of all infertility trials would require all infertility trials to be powered to this end. This would rule out the possibility of smaller, explanatory trials, which may prove useful to the development of interventions. We suggest that procedural outcomes of treatment may be more appropriate for the evaluation of such trials. Live birth could still be reported, if not interpreted, and any intervention should ultimately be tested in confirmatory studies with live birth as the primary outcome. It is worth noting that using live birth as the primary outcome incurs practical disadvantages such as the need for a longer duration of follow up, which delays the release of clinical information and may be problematic in the eyes of funding bodies (Braakhekke *et al.*, 2014a). A compromise might be for journals to allow trial reports to be submitted for peer review with ongoing pregnancy results and, following acceptance of the manuscript, to require study authors to supply live birth results prior to publication. A consensus regarding what should constitute an ongoing pregnancy does not appear to exist at present however. We found a variety of definitions in use, with several studies describing pregnancies as ongoing prior to 12 weeks post transfer, contrary to the definition appearing in IMPRINT (Legro *et al.*, 2014). It was not usual for studies to contain an explicit description of live birth at all, and it was rarer still for studies to include a lower limit of gestation as part of the definition (such as the 20 weeks recommended by IMPRINT) (Legro *et al.*, 2014). Taking live birth as an example, we investigated denominator selection in more detail and found evidence that RCT methodology remains widely misunderstood by researchers and peer reviewers. Of those reporting live birth rates, a third of studies used the subgroup of patients achieving oocyte retrieval or embryo transfer as the denominator, rather than the set of all patients who were randomised earlier in the treatment process. The implications of this analytic strategy are more severe than just a loss of power. Randomised trials represent the gold standard in treatment evaluation due to the fact that random allocation to interventions ensures a balance over confounding factors. When outcomes are reported in subgroups of patients who reached a certain stage of the treatment process, and this does not coincide with the original randomised cohort, the balance is not preserved (Yusuf *et al.*, 1991; Hirji and Fagerland, 2009). Accordingly, any observed differences in outcome may be due to differences in prognostic characteristics rather than treatment effects. The comparative groups are particularly likely to differ when patients with certain characteristics are more or less likely to have a successful stimulation response or to achieve transfer in one arm of the trial (Hirji and Fagerland, 2009). Belief in the existence of such differential effects of treatment is the cornerstone of personalised IVF (Nelson, 2013; Dewailly *et al.*, 2014; La Marca and Sunkara, 2014). We expect that the issue will be more severe the greater the number of participants excluded, although this requires investigation in future simulation studies. The percentage of participants excluded in this sample tended to be less than 10%. A simple strategy to avoid this issue is that used by the Cochrane Gynaecology and Fertility Group, which is to define those participants for whom treatment has failed prior to embryo transfer as having an unsuccessful response. We also note that while it is valid to analyse results per transfer or per oocyte retrieval whenever patients have been randomised at this stage of treatment, reporting outcomes per cycle started may be more relevant to patients deciding whether or not to undertake IVF (Heijnen *et al.*, 2004). It may be argued that pragmatic effectiveness studies should therefore randomise prior to the start of the cycle (Mastenbroek and Repping, 2014). Other examples of statistical naiveté were identified. Some studies reported live birth per embryo transferred, which is problematic since embryos are not statistically independent and the outcome is defined at the level of the patient, rather than of the embryo (Vail and Gardener, 2003). Other studies randomly divided each patient's oocytes or embryos between intervention arms and compared the clinical outcomes between groups of patients who happened to have embryos transferred from only one of the arms. This is not a valid comparison, and may reflect the influence of initiatives promoting the reporting of clinical endpoints in all studies. We also suggest that the tendency to report myriad outcomes carries implications of false effect discovery due to multiple testing and selective emphasis or reporting. In theory, the specification of a primary outcome should offer some protection against these concerns, although in the absence of prospective trial registration there is no guarantee that the primary outcome has been selected in advance (Chan *et al.*, 2004). Moreover, these matters are particularly problematic given the fact that any outcome can be constructed in a variety of ways using the building blocks available combined with the strong emphasis on statistical significance in these trials. Outcome reporting bias would appear to represent an ungovernable potential source of bias in this field given that such a plethora of outcome measures are acceptable to peer reviewers.

Our study has limitations. This review was not comprehensive, as we restricted our sample to English-language publications in peer-reviewed journals. It is not clear however that publication bias represents a concern for a review of outcomes, as the accessibility of any particular study may not be related to the outcome measures used. The subgroup analysis of trials reporting live birth was not prespecified. It should also be noted that the categorisation scheme presented here is entirely arbitrary and was not prospectively designed; another review team likely would have made different decisions relating to the

simplification and presentation of the outcome measures. The exact frequencies we present should be interpreted with this in mind. We believe that our conclusions are not contingent upon our particular scheme. Finally, due to the practice of reporting a trial's results across multiple publications, a small number of included articles may have been secondary publications reporting on particular secondary outcomes. Strictly speaking, the article, rather than the trial, is the unit of analysis in this review. We would suggest that it is appropriate to include these publications, as the decision to exclude them would omit reported outcomes where investigators had split results across two publications.

This is the first review to fully detail the outcomes reported across IVF trials. A previous review restricted their search to highly ranked journals and to studies reporting clinical outcomes (Dapuzzo *et al.*, 2011). This was suitable for the authors' aims of highlighting inconsistency in defining outcomes and underreporting of adverse events. It does not permit the prevalence of each outcome to be calculated however. Additionally, we note that high quality of reporting in all journals, not just the best, is a prerequisite for systematic review, where there is a need to identify all trials (although this would also include older studies, which we have not considered here). A second review found modest rates of reporting of neonatal and maternal outcomes in reproductive medicine trials (Braakhekke *et al.*, 2014b). However, that study restricted focus to outcomes in these two categories and only included trials appearing in Cochrane reviews. Accordingly, the results do not give a complete or representative picture of the current state of outcome reporting in IVF trials. There is massive diversity in numerator and denominator selection in IVF trials due to its multistage nature, and this causes methodological frailty in the evidence base. Existing efforts to improve the situation are certainly useful, although we would urge that future extensions to these projects include guidance on the definition and use of denominators as well as numerators and acknowledge that clinical outcomes may not be appropriate for early phase studies.

Supplementary data

Supplementary data are available at http://humrep.oxfordjournals.org/.

Acknowledgements

We would like to thank James Duffy, Lesley-Anne Carter, Antonia Marsden and two anonymous peer reviewers for helpful comments regarding an earlier draft of the manuscript.

Authors' roles

J.W. designed the study, and undertook the acquisition, analysis and interpretation of data, drafted the manuscript and gave final approval of the version to be published. M.S. designed and conducted the search, contributed to the design of the study, the interpretation of data, drafting and revision of the manuscript and gave final approval of the version to be published. All other authors contributed to the design of the study, the interpretation of data, drafting and revision of the manuscript and gave final approval of the version to be published. J.W. is acting as guarantor for the study.

Funding

J.W. is funded by a Doctoral Research Fellowship from the National Institute for Health Research, supervised by A.V. and S.R.

Conflict of interest

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: J.W. is funded by a Doctoral Research Fellowship from the National Institute for Health Research (DRF-2014-07-050), supervised by A.V. and S.R. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. J.W. declares that publishing research benefits his career. D.R.B. is funded by the NHS as Scientific Director of a clinical IVF service. A.V. and J.W. are statistical editors of the Cochrane Gynaecology and Fertility Group, although the views expressed here do not necessarily represent those of the group; no other relationships or activities that could appear to have influenced the submitted work.

References

Abdalla HI, Bhattacharya S, Khalaf Y. Is meaningful reporting of national IVF outcome data possible? *Hum Reprod* 2010;**25**:9–13.

Blazeby J, Altman DG, Clarke M, Gargon EA, Williamson PR. Core outcome sets and the COMET (core outcome measures in effectiveness trials) initiative; improving the efficiency and value of the research process. *Qual Life Res* 2012;**21**:19–20.

Braakhekke M, Kamphuis EI, Dancet EA, Mol F, van der Veen F, Mol BW. Ongoing pregnancy qualifies best as the primary outcome measure of choice in trials in reproductive medicine: an opinion paper. *Fertil Steril* 2014a;**101**:1203–1204.

Braakhekke M, Kamphuis EI, Van Rumste MM, Mol F, Van Der Veen F, Mol BW. How are neonatal and maternal outcomes reported in randomised controlled trials (RCTs) in reproductive medicine? *Hum Reprod* 2014b; **29**:1211–1217.

Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;**171**:735–740.

Clarke M, Williamson PR. Core outcome sets and systematic reviews. *Syst Rev* 2016;**5**:11.

Dapuzzo L, Seitz FE, Dodson WC, Stetter C, Kunselman AR, Legro RS. Incomplete and inconsistent reporting of maternal and fetal outcomes in infertility treatment trials. *Fertil Steril* 2011;**95**:2527–2530.

Dewailly D, Andersen CY, Balen A, Broekmans F, Dilaver N, Fanchin R, Griesinger G, Kelsey TW, La Marca A, Lambalk C *et al*. The physiology and clinical utility of anti-Mullerian hormone in women (vol 20, pg 370, 2014). *Hum Reprod Update* 2014;**20**:804–804.

Garrido N, Bellver J, Remohi J, Simon C, Pellicer A. Cumulative live-birth rates per total number of embryos needed to reach newborn in consecutive in vitro fertilization (IVF) cycles: a new approach to measuring the likelihood of IVF success. *Fertil Steril* 2011;**96**:40–46.

Germond M, Urner F, Chanson A, Primi M-P, Wirthner D, Senn A. What is the most relevant standard of success in assisted reproduction? The cumulated singleton/twin delivery rates per oocyte pick-up: the CUSIDERA and CUTWIDERA. *Hum Reprod* 2004;**19**:2442–2444.

Griesinger G. Beware of the 'implantation rate'! Why the outcome parameter 'implantation rate' should be abandoned from infertility research. *Hum Reprod* 2016;**31**:249–251.

Heijnen E, Macklon NS, Fauser B. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod* 2004;**19**:1936–1938.

Hirji KF, Fagerland MW. Outcome based subgroup analysis: a neglected concern. *Trials* 2009;**10**:1.

Khan K. The CROWN Initiative: journal editors invite researchers to develop core outcomes in women's health. *BJOG* 2014;**121**:1181–1182.

La Marca A, Sunkara SK. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Hum Reprod Update* 2014;**20**:124–140.

Legro RS, Wu X, Scientific C, Barnhart KT, Farquhar C, Fauser BC, Mol B. Improving the reporting of clinical trials of infertility treatments (IMPRINT): modifying the CONSORT statement. *Hum Reprod* 2014;**29**:2075–2082.

Mastenbroek S, Bossuyt PMM, Heineman MJ, Repping S, van der Veen F. Comment 1 on Staessen *et al.* (2004). Design and analysis of a randomized controlled trial studying preimplantation genetic screening. *Hum Reprod* 2005;**20**:2362–2363.

Mastenbroek S, Repping S. Preimplantation genetic screening: back to the future. *Hum Reprod* 2014;**29**:1846–1850.

Min JK, Breheny SA, MacLachlan V, Healy DL. What is the most relevant standard of success in assisted reproduction? The singleton, term gestation, live birth rate per cycle initiated: the BESST endpoint for assisted reproduction. *Hum Reprod* 2004;**19**:3–7.

Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ* 2001;**165**:1339–1341.

Nelson SM. Biomarkers of ovarian response: current and future applications. *Fertil Steril* 2013;**99**:963–969.

Vail A, Gardener E. Common statistical errors in the design and analysis of subfertility trials. *Hum Reprod* 2003;**18**:1000–1004.

Van Voorhis BJ. Clinical practice. In vitro fertilization. *N Engl J Med* 2007;**356**:379–386.

Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical-trials. *JAMA* 1991;**266**:93–98.